BIOMETRICS 47, 63–76
March 1991

Selecting an Efficient Design for Assessing Exposure–Disease Relationships in an Assembled Cohort

Sholom Wacholder,[1] Mitchell Gail,[1] and David Pee[2]

[1]National Cancer Institute, Biostatistics Branch, 6130 Executive Blvd., EPN/403,
Rockville, Maryland 20892, U.S.A.

and

[2]Information Management Services, Rockville, Maryland 20852, U.S.A.

SUMMARY

We develop approximate methods to compare the efficiencies and to compute the power of alternative potential designs for sampling from a cohort before beginning to collect exposure data. Our methods require only that the cohort be assembled, meaning that the numbers of individuals N_{kj} at risk at pairs of event times t_k and $t_j \geqslant t_k$ are available. To compute N_{kj}, one needs to know the entry, follow-up, censoring, and event history, but not the exposure, for each individual. Our methods apply to any "unbiased control sampling design," in which cases are compared to a random sample of noncases at risk at the time of an event. We apply our methods to approximate the efficiencies of the nested case–control design, the case–cohort design, and an augmented case–cohort design, compared to the full cohort design, in an assembled cohort of 17,633 members of an insurance cooperative who were followed for mortality from prostatic cancer. The assumptions underlying the approximation are that exposure is unrelated both to the hazard of an event and to the hazard for censoring. The approximations performed well in simulations when both assumptions held and when the exposure was moderately related to censoring.

1. Introduction

Retrospective cohort studies can usually be carried out much more rapidly and economically than prospective cohort studies because it is not necessary to wait for the events to occur (Breslow and Day, 1987). Instead, available records are used to define the cohort and to determine for each member the time interval during which he was at risk and when the event of interest, such as diagnosis of cancer, occurred, if it did. We use the term *assembled cohort* to describe a cohort for which this information is available. The purpose of this paper is to show how to compare the efficiencies of various methods of sampling exposure information from an assembled cohort.

To be specific, suppose that age is the time scale that requires tightest control (Breslow et al., 1983) and that the d events occur at distinct ages $t_1 < t_2 < \cdots < t_d$. Define the set R_j of members at risk at t_j and, for $t_k < t_j$, let N_{kj} be the number of individuals who are members of R_k and R_j. We show how the quantities N_{kj} $(k, j = 1, 2, \ldots, d)$ may be used to select an efficient design for sampling subjects to obtain information on an exposure variable X when the hazard of an event at t is $h_0(t)\exp(\beta X)$, X is a scalar, and $h_0(t)$ is the age-specific hazard for a subject with exposure $X = 0$. A more precise definition of an assembled cohort is a cohort for which the quantities N_{kj} have been determined.

Key words: Case–cohort design; Nested case–control design; Proportional hazards; Superpopulation variance; Unbiased control sampling.

One option is to gather exposure information and other covariates on all members of the cohort. However, important cost savings can be achieved with very little loss of efficiency by obtaining exposure information on the d subjects who develop events and on a subset of other cohort members, provided d is small compared to N, the total cohort size, as we shall assume. We denote by C_j the set of all $N_{jj} - 1$ members of R_j who do not have the event at t_j. Liddell, McDonald, and Thomas (1977) proposed taking a small random sample of the members of C_j and comparing their exposures with that of the case at t_j, as in the conditional logistic analysis for time-matched case–control studies (Cox, 1972; Breslow et al., 1978). The procedure has good efficiency relative to the full cohort analysis (Breslow and Patton, 1979; Whittemore and McMillan, 1982) for testing the null hypothesis of no exposure effect ($\beta = 0$). This widely used sampling scheme is termed the *nested case–control design*. Recently, Prentice (1986) proposed a *case–cohort design* in which a simple random sample called the *subcohort* is obtained from the entire cohort, and the exposure of the case at t_j is compared to that of all subcohort members at risk at t_j. The estimation procedure is the same as for a matched case–control analysis, except the "controls" are the subcohort members who are in C_j, the subset of R_j that excludes the case. We also consider an *augmented case–cohort design* in which exposure information is obtained on a random sample from the entire cohort (subcohort S_1) and from a second random sample drawn from all individuals at risk at or beyond a fixed age τ (the subcohort S_2). Prentice (1986) suggested such cohort augmentation but provided no methods for analysis. The difficult variance calculations might be circumvented by means of an extension to the bootstrap procedure given by Wacholder et al. (1989) for the standard case–cohort design. Although the methods for variance estimation have not been worked out in detail, we shall present a simple method to approximate the efficiency of augmented case–cohort designs for an assembled cohort and to determine whether it is even worthwhile to consider such designs for a given application.

The efficiency characteristics of a design depend on the nature of the failure and censoring in the cohort. Robins, Gail, and Lubin (1986) distinguish a *closed* cohort, in which risk sets are nested and monotonically decreasing in time, from an *open* cohort, in which there is no such nesting. In clinical trials, the cohorts are typically closed. The basic time scale is often time after randomization so all members begin to be at risk at $t = 0$. The risk sets are subsequently depleted by competing risks, loss to follow-up, and events of interest but are never supplemented. Many epidemiologic cohorts, on the other hand, are open, because age is used as the basic time scale and subjects' ages vary at the beginning of follow-up. Thus risk sets at later ages may contain subjects who were not at risk at earlier ages.

Self and Prentice (1988) presented general efficiency calculations for the case–cohort design and demonstrated numerically that the design is more efficient than the case–control design for a clinical trial with rare events and no competing risks. Simulations in Wacholder et al. (1989) suggested that in closed cohorts the case–cohort design is more efficient than the case–control design in the absence of competing risks and less efficient with moderate or large competing risks. Langholz and Thomas (1990) presented simulations suggesting that the nested case–control design could be more efficient than the standard case–cohort design for cohorts even with a little censoring. We discuss the relative efficiencies of these two designs and the augmented case–cohort design in the context of a specific assembled cohort of 17,633 men who enrolled in an insurance cooperative in 1965 and who were subsequently followed for mortality from prostatic cancer through 1985 (Bjelke, unpublished Ph.D. thesis, University of Minnesota, 1973). Because subjects had varying ages at enrollment, risk sets at older ages contain some members who had not been at risk at earlier ages. Thus, the assembled cohort we are studying is open.

In Section 2, we define an estimating equation for estimating β and testing $\beta = 0$ for any design with "unbiased control sampling." By *unbiased control sampling*, we mean that the control group, $\tilde{C}_j$, used for comparison with the individual with the event at t_j is a random sample, drawn without replacement from C_j. This class of designs includes all the designs mentioned above. For example, since the subcohort in a case–cohort design is a random sample from the entire cohort, the members in $\tilde{C}_j$ are a random sample from C_j. We present simple approximate variance estimates, called *superpopulation variances*, for the score statistic for testing $\beta = 0$ and for the estimate $\tilde{\beta}$, which are valid when $\beta = 0$ and when censoring is independent of exposure X. These approximate variances can be used to estimate power and to gauge the relative efficiences of these designs using information only on the N_{kj} in the assembled cohort. Section 3 presents efficiency calculations for the insurance cooperative cohort. Section 4 describes the results of simulations to confirm the accuracy of the efficiency calculations derived from the superpopulation variance and to determine whether these calculations can be misleading if censoring depends on X.

2. Methods

2.1 *Analysis via Estimating Equations for Designs with Unbiased Control Sampling*

Let D_j denote the singleton set containing the index of the individual who failed at t_j and let $\tilde{C}_j$ denote the set of indices of the control group, which is a random sample drawn without replacement from C_j. We note that D_j and $\tilde{C}_j$ are disjoint, and we define $\tilde{R}_j = \tilde{C}_j \cup D_j$. As in Prentice (1986), define the logarithm of the pseudo-likelihood by

$$\begin{aligned} l(\beta) &\equiv \sum_j l_j(\beta) \\ &\equiv \Sigma_j\left[\log\{\Sigma I(i \in D_j)\exp(\beta X_i)\} - \log\{\Sigma I(i \in \tilde{R}_j)\exp(\beta X_i)\}\right], \end{aligned} \tag{1}$$

where j indexes event times t_j and i indexes individuals. Here $I(A)$ is an indicator function with value 1 if A is true and 0 elsewhere, sums indexed by j are over all distinct event times, t_j, and other sums are over all individuals in the cohort, unless indicated otherwise. Equation (1) is of the same form as the Cox (1972) partial log-likelihood for the full cohort except that $\tilde{R}_j$ replaces $R_j = C_j \cup D_j$. Letting $U(\beta) = \partial l / \partial\beta$ and $U_j(\beta) = \partial l_j / \partial\beta$, we write the score

$$\begin{aligned} U(\beta) &\equiv \sum_j U_j(\beta) \\ &\equiv \Sigma_j\Big[\Sigma I(i \in D_j)X_i - \{\Sigma I(i \in \tilde{R}_j)\exp(\beta X_i)\}^{-1}\{\Sigma I(i \in \tilde{R}_j)X_i\exp(\beta X_i)\}\Big], \end{aligned} \tag{2}$$

and estimate β by the solution $\tilde{\beta}$ to the equation $U(\beta) = 0$.

A necessary condition for the estimating equation $U(\tilde{\beta}) = 0$ to yield consistent estimates $\tilde{\beta}$ is that $\mathrm{E}\{U(\beta)\} = 0$ (Godambe, 1960). Condition, for the moment, on given R_j, N_{jj}, and n_{jj}. Then the probability that individual i was the member of $\tilde{R}_j$ who had the event in the interval $(t_j, t_j + \Delta]$ is

$$\Delta h_0(t_j)\exp(\beta X_i)\Pr(\tilde{C}_j \mid C_j, i) = \Delta h_0(t_j)\exp(\beta X_i)\binom{N_{jj} - 1}{n_{jj}}^{-1}$$

for a small positive Δ, provided $\tilde{C}_j$ is a random sample from C_j. The corresponding conditional probability of $\tilde{R}_j$ is the sum of such terms over indices i in $\tilde{R}_j$. Therefore,

$$\Pr(i \text{ is the case } \mid \tilde{R}_j, R_j, N_{jj}, n_{jj}) = \exp(\beta X_i)/\Sigma_{i \in \tilde{R}_j}\exp(\beta X_i),$$

which is independent of R_j, N_{jj}, and n_{jj}, where $n_{jj} = \Sigma I(i \in \tilde{C}_j)$. Hence $\Pr(i$ is the case $\mid \tilde{R}_j)$ is given by the expression above and therefore $\mathrm{E}\{U_j(\beta)\} = 0$ and $\mathrm{E}\,U(\beta) = 0$.

As in equation (8) of Prentice (1986), $\text{var}(U) = \Sigma V_{kk} + 2\Sigma_{k<j} V_{kj}$, where $V_{kj} = \text{cov}(U_k, U_j)$, and from Taylor series expansion,

$$\text{var}(\tilde{\beta}) = \left(\sum V_{kk}\right)^{-1}\left\{\sum V_{kk} + 2\sum_{k<j} V_{kj}\right\}\left(\sum V_{kk}\right)^{-1}. \tag{3}$$

Also, the pseudo-score test for $\beta = 0$ is based on $U(0)[\text{var}\{U(0)\}]^{-1/2}$, where V_{kj} are evaluated at $\beta = 0$. Various designs for unbiased control sampling lead to different covariances V_{kj} and different efficiencies. In Section 2.2 we present a simple approximate calculation of $\text{var}\{U(0)\}$ that is useful for assessing the relative efficiencies of alternative designs. We shall assume that the four unbiased sampling plans in Section 2.2 yield consistent asymptotically normal estimates $\tilde{\beta}$ as solutions to equation (2). These properties have been proved for the full cohort analysis (Cox, 1972, 1975; Andersen and Gill, 1982), the nested case–control design (Oakes, 1981), and the case–cohort design (Self and Prentice, 1988), but no such proof has been given for augmented case–cohort sampling.

2.2 *Superpopulation Variance Estimates for the Full-Cohort, Nested Case–Control, Case–Cohort, and Augmented Case–Cohort Designs*

In general, the V_{kk} and V_{kj} depend on β, the nuisance hazard $h_0(t)$, and censoring patterns in a complicated way (Self and Prentice, 1988). However, calculations of these quantities simplify under the assumptions: (A1) $\beta = 0$ and (A2) the mechanism determining when individuals are under observation, such as left truncation and right censoring, is independent of X. Calculations of local efficiency and power for small β may be carried out under these assumptions. However, to the extent that assumption A2 is violated, the following procedures must be regarded as approximate. Simulations suggest that the following methods yield useful results even when assumption A2 is false (§4.3). We call calculations under the assumptions A1 and A2 *superpopulation variance estimates*, because under those assumptions every member of the cohort, whether dead at an early age or alive and at risk at an old age, has an associated X_i that may be regarded as an independent random observation from a superpopulation with mean $\text{E}X = \mu$ and variance $\text{var}(X) = \sigma^2$. Even when it is unreasonable to regard cohort members as a random sample from a superpopulation, these variance estimates are still useful if the cohort is large.

To take advantage of the superpopulation assumptions, we reexpress $U_j(0)$ under $\beta = 0$ as

$$U_j(0) = (n_{jj} + 1)^{-1}\{n_{jj}\Sigma I(i \in D_j)X_i - \Sigma I(i \in \tilde{C}_j)X_i\}. \tag{4}$$

For later use, we define $n_{kj} = \Sigma I(i \in \tilde{C}_k \cap \tilde{C}_j)$ and recall that $N_{kj} = \Sigma I(i \in R_k \cap R_j)$ counts cases in addition to the members of $C_k \cap C_j$. As in Prentice (1986), we shall always let k index the earlier of two event times, $t_k < t_j$. We also define $\delta_{kj} = 1$ if $D_j \subset \tilde{C}_k$ and 0 otherwise.

We call the censoring and survival information needed to calculate N_{kj} and N_{jj} the *assembled cohort history*, ACH, which does not include information on covariates or on who died at t_j. Subsequent designed sampling from the assembled cohort defines the entire survival history for all those who died and all subjects chosen to be in any comparison group, $\tilde{C}_j$. We call this information the *sampling history*, SH. In the sampling designs we consider, all cases are always sampled, so their survival histories are the same in each SH, but the associated quantities δ_{kj} may vary. Conditional on SH, the indicators in equation (4) are fixed constants, and since each index i is associated with an independent, identically distributed observation, X_i, the conditional expectation of $U_j(0)$, given SH, is zero. This follows because exactly n_{jj} individuals contribute to the second sum in (4), conditional on SH. The conditional variance of $U_j(0)$ given

SH is immediately seen to be

$$\operatorname{var}\{U_j(0)\,|\,\mathrm{SH}\} = (n_{jj}+1)^{-2}\{n_{jj}^2\sigma^2 + n_{jj}\sigma^2\}$$

$$= n_{jj}(n_{jj}+1)^{-1}\sigma^2, \qquad (5)$$

since D_j and $\tilde{C}_j$ are disjoint sets of indices. Likewise the conditional covariance is

$$\operatorname{cov}\{U_k(0), U_j(0)\,|\,\mathrm{SH}\} = (n_{kk}+1)^{-1}(n_{jj}+1)^{-1}\{0 + 0 - \sigma^2\delta_{kj}n_{jj} + \sigma^2 n_{kj}\}$$

$$= (n_{kk}+1)^{-1}(n_{jj}+1)^{-1}(n_{kj} - n_{jj}\delta_{kj})\sigma^2. \qquad (6)$$

The two zero terms in equation (6) arise with $t_k < t_j$ because D_k is disjoint from D_j and $\tilde{C}_j$. We comment parenthetically that the calculation $\mathrm{E}\{U_j(0)\,|\,\mathrm{SH}\} = 0$ and calculations leading to (5) and (6) by the superpopulation method are very different from the calculations of Cox (1972) and Prentice (1986), who do not condition on SH but rather take expectations conditional on membership in risk sets R_j. Their approach, unlike the superpopulation calculations, is valid even when assumptions A1 and A2 do not hold.

Under assumptions A1 and A2, we seek $\operatorname{var}\{U_j(0)\,|\,\mathrm{ACH}\}$ and $\operatorname{cov}\{U_k(0), U_j(0)\,|\,\mathrm{ACH}\}$ by averaging quantities (5) and (6) over all realizations consistent with the experimental design for sampling the assembled cohort. Since $\mathrm{E}\{U_j(0)\,|\,\mathrm{SH}\} = 0$, terms like $\operatorname{var}[\mathrm{E}\{U_j(0)\,|\,\mathrm{SH}\}] = \operatorname{var}[0] = 0$ can be ignored. We now calculate $\operatorname{cov}\{U_k(0), U_j(0)\,|\,\mathrm{ACH}\}$ for the full-cohort, nested case–control, case–cohort, and augmented case–cohort designs.

The full-cohort design yields exposure information on everyone in the assembled cohort, so sets $\tilde{C}_j = R_j - D_j = C_j$. Since ACH determines the quantities N_{kj}, the terms $n_{jj} = N_{jj} - 1$ in (5) are fixed, and $\operatorname{var}\{U_j(0)\,|\,\mathrm{ACH}\} = (N_{jj} - 1)\sigma^2/N_{jj}$ as in Cox (1972), except that σ^2 replaces the sample variance of the exposures of members of R_j. In (6), $n_{jj} = N_{jj} - 1$ is fixed, but $n_{kj} = N_{kj} - \delta_{kj}$ varies. The expectation

$$\mathrm{E}(\delta_{kj}\,|\,\mathrm{ACH}) = N\Pr(i \in C_k)\Pr(i \in \tilde{C}_k \,|\, i \in C_k)$$

$$\times \Pr(i \in R_j \,|\, i \in \tilde{C}_k \cap C_k)\Pr(i \in D_j \,|\, i \in R_j \cap \tilde{C}_k \cap C_k)$$

$$= N\{(N_{kk} - 1)/N\}\{1\}\{N_{kj}/(N_{kk} - 1)\}\{1/N_{jj}\}$$

$$= N_{kj}/N_{jj},$$

since $\Pr(i \in D_j \,|\, i \in R_j \cap \tilde{C}_k \cap C_k) = \Pr(i \in D_j \,|\, R_j)$ by assumptions A1 and A2. Thus the expectation of (6) given ACH is zero, regardless of whether the assembled cohort is open or closed.

For nested case–control sampling with a fixed number of noncases, l_j, chosen at t_j, the quantities $n_{jj} = l_j$ in (4) are constant, and (5) yields standard variances for matched case–control studies (Ury, 1975). In (6),

$$\mathrm{E}(n_{kj}\,|\,\mathrm{ACH}) = N\{(N_{kk} - 1)/N\}\{l_k/(N_{kk} - 1)\}$$

$$\times\{N_{kj}/(N_{kk} - 1)\}\{(N_{jj} - 1)/N_{jj}\}\{l_j/(N_{jj} - 1)\}$$

$$= l_k l_j N_{kj}/\{N_{jj}(N_{kk} - 1)\},$$

from the laws of conditional probabilities used above. Likewise,

$$\mathrm{E}(\delta_{kj}\,|\,\mathrm{ACH}) = N\{(N_{kk} - 1)/N\}\{l_k/(N_{kk} - 1)\}\{N_{kj}/(N_{kk} - 1)\}\{1/N_{jj}\}$$

$$= l_k N_{kj}/\{N_{jj}(N_{kk} - 1)\}.$$

Hence (6) vanishes in expectation given ACH for the nested case–control design.

For the case–cohort design (Prentice, 1986) and other more general unbiased control sampling plans (§1), one cannot easily compute the exact expectation of (5) or (6) given ACH because quantities like $(n_{jj} + 1)$ in the denominators vary. Nonetheless, a reasonable approximation, especially for large n_{kj}, is obtained by replacing each of the random quantities n_{kj} and δ_{kj} in (5) and (6) by its expectation. In particular,

$$\begin{aligned}\mathrm{E}(n_{jj}\,|\,\mathrm{ACH}) &= N_{jj}\,\Pr(i \in C_j \,|\, i \in R_j)\Pr(i \in \tilde{C}_j \,|\, i \in C_j \cap R_j)\\ &= N_{jj}\{(N_{jj} - 1)/N_{jj}\}\rho_j\\ &= (N_{jj} - 1)\rho_j,\end{aligned}$$

where

$$\rho_j \equiv \Pr(i \in \tilde{C}_j \,|\, i \in C_j \cap R_j) = \Pr(i \in \tilde{C}_j \,|\, i \in C_j).$$

Likewise, $\mathrm{E}(n_{kk}\,|\,\mathrm{ACH}) = (N_{kk} - 1)\rho_k$, and

$$\begin{aligned}\mathrm{E}(n_{kj}\,|\,\mathrm{ACH}) &= N_{kj}\,\Pr(i \in C_k \cap C_j \,|\, i \in R_k \cap R_j)\Pr(i \in \tilde{C}_k \cap \tilde{C}_j \,|\, i \in C_k \cap C_j \cap R_k \cap R_j)\\ &= N_{kj}\,\Pr(i \in C_j \,|\, i \in R_k \cap R_j)\rho_{kj}\\ &= N_{kj}\{(N_{jj} - 1)/N_{jj}\}\rho_{kj},\end{aligned}$$

where

$$\rho_{kj} \equiv \Pr(i \in \tilde{C}_k \cap \tilde{C}_j \,|\, i \in C_k \cap C_j \cap R_k \cap R_j) = \Pr(i \in \tilde{C}_k \cap \tilde{C}_j \,|\, i \in C_k \cap C_j).$$

Finally,

$$\begin{aligned}\mathrm{E}(\delta_{kj}\,|\,\mathrm{ACH}) &= N\,\Pr(i \in \tilde{C}_k \cap D_j \,|\, \mathrm{ACH})\\ &= N\,\Pr(i \in R_k \cap R_j)\Pr(i \in C_k \,|\, R_k \cap R_j)\\ &\quad \times \Pr(i \in \tilde{C}_k \,|\, i \in C_k \cap R_k \cap R_j)\Pr(i \in D_j \,|\, i \in \tilde{C}_k \cap C_k \cap R_k \cap R_j)\\ &= N(N_{kj}/N)(1)(\rho_k)(1/N_{jj}) = \rho_k N_{kj}/N_{jj}.\end{aligned}$$

Substituting these results in (5) and (6), we obtain the approximations

$$V_{jj} \equiv \mathrm{var}\{U_j(0)\,|\,\mathrm{ACH}\} \approx \sigma^2\{(N_{jj} - 1)\rho_j\}\{(N_{jj} - 1)\rho_j + 1\}^{-1} \tag{7}$$

and

$$\begin{aligned}V_{kj} \equiv \mathrm{cov}\{U_k(0), U_j(0)\,|\,\mathrm{ACH}\} &\approx \sigma^2\{N_{kj}(N_{jj} - 1)/N_{jj}\}\{\rho_{kj} - \rho_k\rho_j\}\\ &\quad \times \{(N_{kk} - 1)\rho_k + 1\}^{-1}\{(N_{jj} - 1)\rho_j + 1\}^{-1}.\end{aligned} \tag{8}$$

In particular, note that for the full cohort design $\rho_{kj} = \rho_k = \rho_j = 1$ so that (8) vanishes, just as in the exact calculation above. Likewise, for nested case–control sampling in which controls at t_j are properly selected (Robins et al., 1986) without replacement from C_j and independently of other selections, $\rho_{kj} = \rho_k\rho_j$, so that (8) vanishes, just as in the exact calculation above for 1-to-l case–control sampling with $\rho_k = l/(N_{kk} - 1)$. We now apply the results (7) and (8) to the case–cohort and augmented case–cohort designs.

In the case–cohort design (Prentice, 1986), we select a "subcohort" of fN individuals at random from the entire cohort of N subjects. The sets $\tilde{C}_j$ consist of those members of C_j who are also in the subcohort. Hence $\rho_{kj} = \rho_k = \rho_j = f$, and from (8) we find that covariances are positive and proportional to $f - f^2 = f(1 - f)$. In particular, covariances vanish as the sampling fraction f tends to 0 or as f tends to 1, the full-cohort design.

Table 1

Approximate efficiencies for estimating β compared to the full-cohort design[a]

	Control-to-case ratio, m				
	1	3	5	10	20
Case–control	.500	.750	.834	.909	.953
Case–cohort	.370	.630	.739	.850	.921
Optimal[b] augmented case–cohort	.404	.667	.769	.872	.934

[a] Efficiencies are calculated from the superpopulation model as in Section 2.2.

[b] A grid search was performed on f_1, f_2, and τ, subject to the constraint $f_1 N + f_2 N_\tau = 55m$, by using values $f_1 N/(f_1 N + f_2 N_\tau) = .1, .2, \ldots, .9$ and by allowing τ to be the death times corresponding to events $[5.5i]$ for $i = 1, 2, \ldots, 9$, where $[\cdot]$ denotes the next lower integer. The smallest variance thus found for the estimate of β was used to compute efficiency. For the case $m = 5$, the optimizing values were $f_1/(f_1 + f_2) = .5$ and τ was the 17th death time.

We define the augmented case–cohort design as follows. First sample a fraction f_1 from the entire cohort and call these $f_1 N$ individuals SC_1. Then, define an augmentation time τ, and independently sample a fraction f_2 from the set $R_\tau - SC_1$, where R_τ is the set of all N_τ persons at risk at or beyond τ. The sets $\tilde{C}_j$ consist only of members of $SC_1 \cap C_j$ for $t_j < \tau$ and of members of $(SC_1 \cup SC_2) \cap C_j$ otherwise. For $t_j < \tau$, $\rho_j = f_1$, whereas, for $t_j \geqslant \tau$, $\rho_j = f_1 + f_2(1 - f_1)$. If $t_k < \tau$, $\rho_{kj} = f_1$. If $t_k \geqslant \tau$, $\rho_{kj} = f_1 + f_2(1 - f_1)$. These values of ρ_j and ρ_{kj} may be substituted into (7) and (8) to obtain needed variances and covariances.

The augmented case–cohort design depends on τ, the point that determines who is eligible for the second subcohort, as well as the sampling fractions f_1 and f_2. In the following comparisons among designs, we perform a grid search on τ, f_1, and f_2 to find the values τ^*, f_1^*, and f_2^* that minimize var($\tilde{\beta}$) subject to the constraint that $f_1 N + f_2 N_\tau$ equals the fixed total number of subjects in the two subcohorts. We compare the "optimal" augmented case–cohort design with other designs in Table 1, but in simulation studies (§4) we retain the optimal design from the original assembled cohort rather than reoptimize for each simulated repetition.

3. Example

3.1 *Calculation of Superpopulation Variances and Corresponding Efficiencies for the Insurance Cooperative Assembled Cohort*

The insurance cooperative cohort contained 17,633 men whose ages at entry varied from 30 to 97. During 20 years of follow-up, 55 deaths from prostatic cancer were identified. From each individual's data on times of entry, loss to follow-up, and time of death, the quantities N_{kj} were calculated for $k = 1, 2, \ldots, 55$; $j = k, k + 1, \ldots, 55$. This assembled cohort was "open" on the time scale defined by age because individuals entered at different ages; the individual data may be described as variably left-truncated and right-censored since ages at entry and end of follow-up varied.

From equations (3) and (7) with $V_{kj} = 0$, the variance of $\tilde{\beta}_{\mathrm{FC}}$, the estimator from the full cohort, was found to be var($\tilde{\beta}_{\mathrm{FC}}$) = $.0182/\sigma^2$, where σ^2 is the unknown variance of exposures in the population. Because N_{jj} is always large, V_{jj} in (7) is very nearly σ^2, so that var($\tilde{\beta}_{\mathrm{FC}}$) $\approx (\sigma^2 55)^{-1}$, a result found in Cox (1972). Indeed, to three significant figures, $1/55 = .0182$ agrees with the result derived from (7). The unknown σ^2 cancels from efficiencies, which are the ratios of var($\tilde{\beta}_{\mathrm{FC}}$) to variances of estimates of β from other designs (Table 1).

With m controls for each case, the variance ratios computed from (3), (7), and (8) for the nested case–control design are virtually identical (Table 1) to the well-known efficiency formula $m/(m + 1)$ given by Ury (1975). This result follows by noting that $(N_{jj} - 1)\rho_k$ is very nearly

m in equation (7) and that (8) vanishes. The approximate efficiency of the case–cohort design is less than that of the case–control design for this assembled cohort (Table 1). The case–cohort sampling fraction $f = 55m/17{,}633$ was used (§2.2). With $m = 1$, the case–cohort design is only 74% as efficient as the case–control design (Table 1). These efficiency calculations ignore the fact that slightly fewer than $(m + 1)d$ distinct cases and controls are utilized in these designs, because of potential overlap among controls and between cases and controls. For example, the case–cohort design typically involves $55m[1 - 55/17{,}633] = 54.8m$ rather than $55m$ noncases. Likewise, Langholz and Thomas (1990) discuss overlap in the case–control design. In our example, corrections for overlap do not affect efficiency calculations appreciably. However, if many events are associated with small overlapping risk sets, R_j, such corrections may be required.

Augmenting the case–cohort design with a second subcohort produces only a minor improvement in efficiency, compared with the case–cohort design, and in no case is the augmented case–cohort design more efficient than the case–control design for this assembled cohort (Table 1).

Perhaps the most important conclusion from these calculations is that for $m \geqslant 5$, all three economical designs vastly reduce the number of subjects for whom full covariate information is required at little cost in statistical efficiency. Differences in efficiency among the three economical designs are small for $m \geqslant 5$.

4. Simulation Study of Superpopulation Variance

4.1 *Design of Simulations*

We sought to investigate the properties of the superpopulation variance via simulation both when assumptions A1 and A2 were satisfied and in special simulations where the assumption A2 was purposely violated. We performed eight simulation experiments, each consisting of the following steps:

1. Randomly resample with replacement a cohort of 17,633 from the original cohort of size 17,633, using the original entry, censoring, or event times, and failure indicator for each individual in the insurance cohort (Bjelke, unpublished Ph.D. thesis, University of Minnesota, 1973).
2. Subjects in the resampled cohort are independently assigned to one of two "factories" with probability $\frac{1}{2}$.
3. Exposure is randomly assigned to subjects according to separate exposure distributions in factory A and factory B. Within factories, exposures were independently distributed.

If only these steps are performed, assumptions A1 and A2 are satisfied. To test the effects of violations of A2, we added one additional step in some of the experiments:

4. All subjects assigned to factory B are censored at age 67.9, which is the median age at death from prostatic cancer.

If the exposure distributions are different in factories A and B, step 4 induces a violation of assumption A2. Assumption A1 is not violated by step 4 since exposure is unrelated to outcome both before and after age 67.9. The expected number of events when step 4 is included is $(\frac{3}{4})(55) = 41.25$.

In each simulation the matrix N_{kj} changed, and we recalculated the superpopulation variances. We also estimated the log-hazard ratio from the full-cohort analysis, from a case–control sample with $m = 5$ controls per case, from a case–cohort sample with subcohort size md, and from an augmented case–cohort sample with a total of md subjects in the two subcohorts. In the simulations, d was a random variable with mean 55, except in studies which included step 4, for which the mean was 41.25. Ties were broken at random.

Table 2
Designs of simulation experiments

Experiment #	Additional censoring in factory B?	Exposure distribution in factory A	Exposure distribution in factory B	A2 violated?
1	No	Bern(.5)[a]	Bern(.5)	No
2	Yes	Bern(.5)	Bern(.5)	No
3	No	$N(0, 1)$	$N(0, 1)$	No
4	Yes	$N(0, 1)$	$N(0, 1)$	No
5	Yes	Bern(.25)	Bern(.75)	Yes
6	Yes	Bern(.05)	Bern(.95)	Yes
7	Yes	$N(0, 1)$	$N(1.15, 1)$	Yes
8	Yes	$N(0, 1)$	$N(4, 1)$	Yes

[a] Bern(p) stands for the Bernoulli distribution with probability p, and $N(\mu, \sigma^2)$ stands for the normal distribution with mean μ and variance σ^2.

Other details of the simulation design are in Section 4.1. A1 is not violated for any of the studies.

Since the N_{kj} varied with each resampling, the optimal augmented case–cohort design also varied. Rather than optimize the cutpoint for each sampling, we kept f_1 and f_2 from the original cohort, and chose τ as the quantile of the death time distribution that was "optimal" in the original assembled cohort.

Assumption A2 is violated in experiments 5, 6, 7, and 8 (Table 2) but not in experiments 1, 2, 3, and 4, because although there is differential censoring in experiments 2 and 4, the distribution of exposures is the same in each factory in those experiments. The means of 0.0 and 1.15 in experiment 7 and 0.0 and 4.0 in experiment 8 were chosen so that the standardized differences in means would nearly equal those in experiments 5 and 6, respectively. Assumption A1 is never violated.

4.2 *Results of Simulations to Verify the Accuracy of the Superpopulation Variance Estimates and Efficiency Approximations When Assumptions* A1 *and* A2 *Hold*

Superpopulation assumptions A1 and A2 hold for studies 1, 2, 3, and 4 (Table 3). The average estimated β was close to zero for all designs in all studies, and in no case was there statistically significant evidence against H_0: $\beta = 0$ at the .05 level (data not shown). If the average superpopulation variance were the true variance of $\tilde{\beta}$, then the ratio of the empirical (sample) variance of $\tilde{\beta}$ to the average superpopulation variance would fall within the interval (.914, 1.075) with probability .95 (see Beyer, 1968, Table V.2). Except for experiment 4, the ratios of empirical variances to average superpopulation variance were within 6% of 1.0, and even in experiment 4 the discrepancies were modest. The empirical efficiencies, which were estimated as ratios of the empirical variance of $\tilde{\beta}$ for the full-cohort design to the empirical variance of $\tilde{\beta}$ for an alternative design, agreed with the average ratio of the respective superpopulation variances even more closely and support the contention that case-control sampling is slightly more efficient than the two case-cohort designs. Thus the superpopulation estimates of variances and efficiencies performed well for this assembled cohort in four different conditions of exposure and censoring.

4.3 *Results of Simulations to Test Robustness to Violations of Assumption* A2

Moderate violations of assumption A2 caused the superpopulation variance approximations to underestimate the empirical variances of $\tilde{\beta}$ by up to 18% (experiments 5 and 7, Table 4), but efficiency estimates based on the superpopulation model continued to yield reliable guidance on the relative performance on case–control and case–cohort designs. More extreme violations of assumption A2 yielded large discrepancies between the empirical variances and the variances predicted from the superpopulation model (experiments 6 and 8, Table 4), and, in these cases, the superpopulation model suggested that the case–control design would be more efficient,

Table 3

Superpopulation and empirical variances and variance ratios when assumptions A1 *and* A2 *hold*[a]

		Variance				Efficiency[b]		
Experiment #	Design	Average superpopulation	Empirical	Average Wald[c]	Superpopulation to empirical ratio	Average superpopulation	Empirical	Superpopulation to empirical ratio
1	Full cohort	.0743	.0746	.0757	.99	—	—	—
	Case-control	.0891	.0880	.0905	1.01	.834	.848	.98
	Case-cohort	.1014	.0991	—	1.02	.732	.753	.97
	Augmented case-cohort	.0977	.0996	—	.98	.759	.749	1.01
2	Full cohort	.0990	.103	.102	.96	—	—	—
	Case-control	.119	.125	.122	.95	.834	.830	1.00
	Case-cohort	.137	.145	—	.95	.722	.715	1.01
	Augmented case-cohort	.134	.137	—	.97	.738	.753	.98
3	Full cohort	.0185	.0181	.0185	1.03	—	—	—
	Case-control	.0222	.0224	.0227	.99	.834	.807	1.03
	Case-cohort	.0252	.0255	—	.99	.733	.709	1.03
	Augmented case-cohort	.0243	.0237	—	1.03	.759	.763	1.00
4	Full cohort	.0250	.0256	.0250	.98	—	—	—
	Case-control	.0299	.0320	.0307	.93	.834	.799	1.04
	Case-cohort	.0345	.0374	—	.93	.721	.685	1.05
	Augmented case-cohort	.0338	.0387	—	.87	.736	.661	1.11

[a] Based on 1,000 resamplings from the study cohort, as described in the text.

[b] Ratio of variance of $\tilde{\beta}$ from full cohort to variance of $\tilde{\beta}$ estimated from alternative design.

[c] The Wald variance is $\{-\ddot{l}(\tilde{\beta})\}^{-1}$, where $-\ddot{l}(\tilde{\beta})$ is the observed information from the partial likelihood.

Table 4
Superpopulation and empirical variance and variance ratios when assumption A2 *is violated*[a]

Experiment #	Design	Variance				Efficiency[b]		
		Average superpopulation	Empirical	Average Wald[c]	Superpopulation to empirical ratio	Average superpopulation	Empirical	Superpopulation to empirical ratio
5	Full cohort	.0994	.117	.111	.85	—	—	—
	Case-control	.119	.136	.133	.88	.834	.866	.96
	Case-cohort	.137	.156	—	.88	.723	.753	.96
	Augmented case-cohort	.134	.159	—	.85	.737	.739	1.00
6	Full cohort	.099	.132	.139	.75	—	—	—
	Case-control	.118	.158	.168	.75	.834	.834	1.00
	Case-cohort	.137	.162	—	.85	.721	.818	.88
	Augmented case-cohort	.134	.168	—	.79	.736	.786	.94
7	Full cohort	.0185	.0194	.0202	.95	—	—	—
	Case-control	.0222	.0238	.0248	.93	.834	.818	1.02
	Case-cohort	.0256	.0271	—	.94	.721	.716	1.01
	Augmented case-cohort	.0251	.0289	—	.87	.736	.672	1.10
8	Full cohort	.0049	.0075	.0068	.66	—	—	—
	Case-control	.0059	.0090	.0082	.66	.834	.833	1.00
	Case-cohort	.0068	.0094	—	.72	.724	.793	.91
	Augmented case-cohort	.0066	.0091	—	.73	.738	.820	.90

[a] Based on 1,000 resamplings from the study cohort, as described in the text.

[b] Ratio of variance of $\tilde{\beta}$ from full cohort to variance of $\tilde{\beta}$ estimated from alternative design.

[c] The Wald variance is $\{-\ddot{l}(\tilde{\beta})\}^{-1}$, where $-\ddot{l}(\tilde{\beta})$ is the observed information from the partial likelihood.

whereas the empirical results indicate that the three designs have nearly the same efficiency. We conclude that the superpopulation model is useful for selecting a design in the presence of moderate, but not severe, violations of assumption A2 and that sample size calculations, which depend on good estimates of variance, can be seriously distorted by violations of A2.

5. Discussion

We have shown how to use the superpopulation model to compare the efficiencies of a variety of potential designs for unbiased control sampling from an assembled cohort. These methods may be used not only for the four designs studied in this paper, but also for other unbiased control sampling designs for which analytical methods for variance estimation have yet to be developed. These calculations depend only on the numbers N_{kj} at risk in the assembled cohort and seem to be robust to moderate associations between exposure and follow-up pattern, i.e., violation of assumption A2. By carrying out such calculations within separate strata, these methods can be extended to efficiency calculations for stratified analyses. An alternative approach to assessment of efficiencies is to resample from the original cohort as in Section 4.1 and to compare empirical variances of estimates of β obtained from various designs. Halpern and Brown (1987) likewise exploit simulation methods for designing complex clinical trials.

While these calculations are useful for selecting a sampling design, they should not be used for analyzing the final data. For this purpose, variance estimates that are valid regardless of β or censoring pattern should be used, as given, for example, by Cox (1972) for the full-cohort design, by Breslow and Day (1980, Chap. 7) for the nested case-control design, and by Prentice (1986), Self and Prentice (1988), and Wacholder et al. (1989) for the case-cohort design.

We have emphasized efficiency calculations for an assembled cohort with fixed N_{kj}. Wacholder et al. (1989) have applied similar ideas to planning clinical trials in which the N_{kj} are not yet available by replacing N_{kj} by estimated values of N_{kj}. These methods are useful for complex accrual and censoring patterns. Self and Prentice (1988) give a more comprehensive theoretical treatment of such power calculations. These studies indicate that the case-cohort design can be slightly more efficient than the case-control design in closed cohorts with little censoring, whereas the case-control design can be more efficient with heavy censoring.

Our efficiency calculations and simulations were based on a single large cohort. Further studies would be desirable to document the validity and robustness of the procedures we describe.

Each of the three alternatives to the full-cohort analysis drastically reduced sample size requirements at little cost in statistical efficiency. The benefits of augmented case-cohort sampling were modest, compared to the simple case-cohort design (Tables 1, 3, and 4). It is possible that complete reoptimization of the augmented case-cohort sampling for each simulation might have improved the performance of this design slightly, compared with results in Tables 3 and 4, but data in Table 1 suggest that even completely optimized case-cohort augmentation will not perform as well as case-control sampling for this cohort.

One might expect the efficiency of the two case-cohort designs to improve if the exposure X were a time-dependent covariate, rather than a fixed covariate. For example, if X were the cumulative exposure to radiation, then the covariances between pairs of scores $U_k(0)$ and $U_j(0)$ would be smaller than shown in (6), and the efficiency of $\tilde{\beta}$ from the case-cohort designs would be increased. However, extra cost might be incurred if more exposure data or repeated measurements of a covariate were required.

Although the various unbiased sampling schemes we studied have slight differences in efficiency, the main observation from this work is that each of these designs can produce important cost savings. Thus the choice among such designs should often be made on the basis of ease of implementation and other practical considerations. An important practical advantage of the case-cohort design is that the same controls may be used to study several diseases (see

Kupper, McMichael, and Spirtas, 1975; Prentice, 1986). Nonetheless, if one has access to a large assembled cohort, it seems well worthwhile to calculate N_{kj} and to compare the efficiencies of alternative designs.

Résumé

Nous avons développé des méthodes d'approximation afin de comparer l'efficience et de calculer la puissance des différents plans d'échantillonage possibles à partir d'une cohorte, avant de commencer à recueillir des données sur l'exposition. Ces méthodes nécessitent seulement une cohorte dite "rassemblée", c'est à dire où les nombres N_{kj} de sujets à risque pour chaque couple de temps correspondant à des événements t_k et $t_j \geqslant t_k$ sont connus. Pour calculer N_{kj}, il faut connaître, pour chaque sujet, la date d'entrée, le suivi, la notion de censure ou d'événement, sans avior besoin de la notion d'exposition. Nos méthodes s'appliquent à "tout plan d'échantillonnage non biaisé à partir d'une population témoin," où les cas sont comparés à un échantillon tiré au sort de témoins à risque. Nous avons testé nos méthodes afin d'estimer l'efficience d'une étude cas-témoins emboitée, d'une étude cas-cohorte, ou d'une étude de cohorte augmentée, comparée à une enquête de cohorte exhaustive, à partir des données provenant d'une cohorte "rassemblée" de 17633 members d'une société d'assurance suivis pour étudier la mortalité par cancer de la prostate. Notre approximation suppose une exposition indépendante à la fois du risque d'apparition d'un événement et du temps de censure. Les approximations se sont révélées satisfaisantes au cours des simulations quand les hypothèses étaient vérifiées et lorsqu'il existait une liaison modérée entre l'exposition et le processus de censure.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large-sample study. *The Annals of Statistics* **10**, 1100–1120.

Beyer, W. H. (1968). *CRC Handbook of Tables for Probability and Statistics*. Cleveland, Ohio: Chemical Rubber Company.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research, Vol. 1. The Analysis of Case-Control Studies* (IARC Scientific Publications No. 32). Lyon: International Agency for Research on Cancer.

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research, Vol. 2. The Design and Analysis of Cohort Studies* (IARC Scientific Publications No. 82). Oxford: Oxford University Press.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L., and Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology* **108**, 299–307.

Breslow, N. E., Lubin, J. H., Markek, P., and Langholz, B. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association* **78**, 1–12.

Breslow, N. E. and Patton, J. (1979). Case-control analysis of cohort studies. In *Energy and Health*, N. E. Breslow and A. S. Whittemore (eds), 226–242. Philadelphia: Society of Industrial and Applied Mathematics.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Godambe, V. P. (1960). An optimal property of regular maximum likelihood estimation. *Annals of Statistics* **31**, 1208–1211.

Halpern, J. and Brown, W. B. (1987). Designing clinical trials with arbitrary specification of survival functions and for the log rank or generalized Wilcoxon test. *Controlled Clinical Trials* **8**, 177–189.

Kupper, L. L., McMichael, A. J., and Spirtas, R. (1975). A hybrid epidemiologic study design useful in estimating relative risk. *Journal of the American Statistical Association* **70**, 524–528.

Langholz, B. and Thomas, D. C. (1990). Nested case-control and case-cohort methods of sampling from a cohort: A critical comparison. *American Journal of Epidemiology* **131**, 169–176.

Liddell, F. D. K., McDonald, J. C., and Thomas, D. C. (1977). Methods of cohort analysis: Appraisal by application to asbestos mining (with Discussion). *Journal of the Royal Statistical Society, Series A* **140**, 469–491.

Oakes, D. (1981). Survival times: Aspects of partial likelihood (with Discussion). *International Statistical Review* **49**, 235–264.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–12.

Robins, J., Gail, M. H., and Lubin, J. H. (1986). More on biased selection of controls. *Biometrics* **42**, 293–299.

Self, S. G. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics* **16**, 64–81.

Ury, H. K. (1975). Efficiency of case-control studies with multiple controls per case: Continuous or dichotomous data. *Biometrics* **31**, 643–649.

Wacholder, S., Gail, M. H., Pee, D., and Brookmeyer, R. (1989). Alternative variance and efficiency calculations for the case-cohort design. *Biometrika* **76**, 117–123.

Whittemore, A. S. and McMillan, A. (1982). Analyzing occupational cohort data: Application to uranium miners. In *Environmental Epidemiology: Risk Assessment*, R. L. Prentice and A. S. Whittemore (eds), 65–81. Philadelphia: Society of Industrial and Applied Mathematics.

Received July 1989; *revised February* 1990; *accepted March* 1990.